Whither Speech Recognition: The Next 25 Years

In 25 years, speech recognition has evolved from a "futile" endeavor to commerical reality.

David B. Roe and Jay G. Wilpon

A quarter century ago, the influential John Pierce wrote an article questioning the prospects of speech recognition and disparaging the "mad inventors and unreliable engineers" working in the field [1]. Yet today, speech recognition is widely used in several applications, especially telecommunications systems. Cellular telephones use single-chip processors to recognize spoken commands, and multichannel voice recognition systems automate the process of setting up long distance telephone calls. Personal computers understand speech from physically impaired people. The success of the commercial applications in the United States, Canada, and Japan is inarguable. Meanwhile, in laboratories throughout the world, experimental systems recognize fluent sentences constructed from vocabularies of thousands of words. Accuracy in some experimental systems is better than 95 percent.

How can we explain the advance in this technology in these 25 years? Was Pierce's prediction overly pessimistic? Or have there been breakthroughs in fundamental science that could not have been anticipated? What has happened in the last 25 years to bring about this success?

In his article, "Whither Speech Recognition?" Pierce argued that attempting to build a speech-recognition system was futile because the task of speech understanding is too difficult for any machine. Such a speech understanding system would require tremendous advances in linguistics, natural language, and knowledge of everyday human experience. In this prediction he was completely correct: there is still no speech recognizer that can transcribe natural speech as well as a trained stenographer, because no machine has the required knowledge and experience of human language. Furthermore, this ultimate goal is still not within sight in 1993. Pierce went on to describe the motivation for speech recognition research: "The attraction [of speech recognition] is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon." His influential article was successful in curtailing, but not stopping, speech recognition research.

What Pierce's article failed to foretell was that even limited success in speech recognition — simple, small-vocabulary speech recognizers — would have interesting and important applications. In 1980 George Doddington, in another "Whither Speech Recognition?" article [2], pointed this out. He emphasized that it was unnecessary to build the ultimate speech understanding system with full human capabilities to get simple information over the telephone, or to give commands to personal computers. In the decade since Doddington's article, tens of thousands of these "limited" speech recognition systems have been put into use, and we now see the beginnings of a speech recognition industry as the technology becomes commercially available [3].

In this article, we present an overview of the science of speech recognition, and discuss some of the ways that people are making use of speech recognition technology. We also bravely (or foolishly) attempt to predict the future prospects of speech recognition. We do not claim any more wisdom than Pierce or Doddington, but we have the benefit of a quarter century of speech recognition history from which to try to project into the next 25 years.

Fundamentals of Speech Recognition

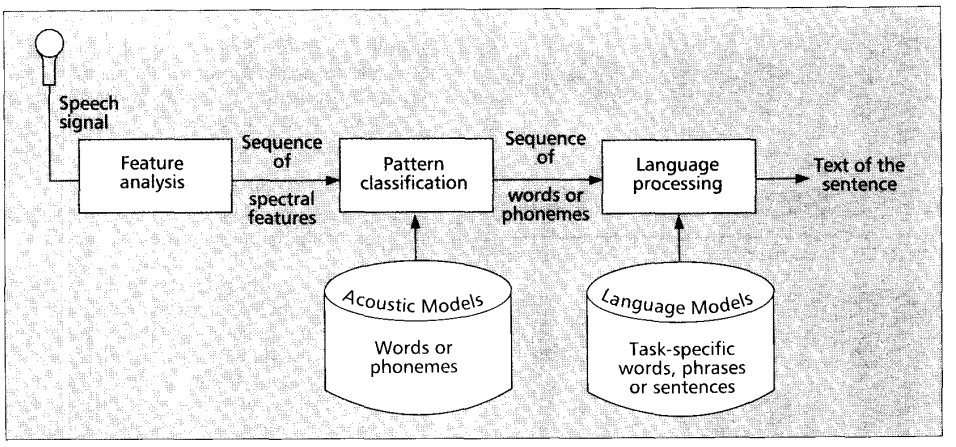
The field of speech recognition is based upon our knowledge of speech and spoken language [4]. Our understanding of speech has grown enormously in the past two decades [5, 6]. The related disciplines of speech coding [7] and text-to-speech synthesis [8] have contributed to our understanding of human-machine communications by voice [9].

Speech recognition is fundamentally a pattern classification task. The objective is to take an input pattern, the speech signal, and classify it as

DAVID B. ROE is a supervisor in the Speech Research Department at AT&T Bell Laboratories.

JAY G. WILPON is a distinguished member of the technical staff at AT&T Bell Laboratories.

0163-6804/93/$03.00 © 1993 IEEE



■ **Figure 1.** *The process of speech recognition.*

a sequence of stored patterns that have previously been learned (Fig. 1). These stored patterns may be made up of units we call words, or shorter units called phonemes (the smallest contrastive sounds of a language). If these patterns were invariant and unchanging, the problem would be trivial: simply compare sequences of features with the stored patterns, and find exact matches when they occur.

The fundamental difficulty of speech recognition is that the speech signal is highly variable due to different speakers, different speaking rates, different contexts, and different acoustic conditions [4]. The task is to determine which of the variations in speech are relevant to speech recognition (for instance, the distinction in pronunciation between "father" and "farther") and which variations are not relevant (for instance, a New England accent). Even with powerful statistical techniques, it is not yet clear how to automatically generalize, from examples of speech, which of the variations in speech are significant and which are not.

The study of speech recognition is based on three principles [10]. First, that the information in the speech signal can be accurately represented by the short-term amplitude spectrum of the speech waveform. This allows us to extract features based on the short-term amplitude spexctrum from speech, and to confidently use these features as the basis for pattern matching. Second, the contents of the speech signal can be expressed in written form. Furthermore, the meaning of speech can be written as a sequence of a few dozen symbols selected from the characters in an alphabet, or from the phonetic symbols in a lexicon. This principle gives us confidence that the meaning of an utterance is preserved as we transcribe a sequence of acoustic features to a sequence of phonetic symbols. Third, speech recognition is a cognitive process. In human speech understanding it is impossible to separate perception of sounds from the grammatical, semantic, and pragmatic structures of language. Because spoken language is meaningful, both semantic and pragmatic information can be valuable guides to speech recognition when acoustic information alone is ambiguous. Unfortunately, current speech recognizers, in their simplicity, have taken little advantage of semantic and pragmatic structures except in small, constrained tasks.

The field of speech recognition is multidisciplinary, and knowledge from a variety of fields continues to be essential to progress in this area:

- Physics (acoustics): the science of room reverberation, the study of microphone design, the acoustics of the human vocal tract, and the mechanics of the human ear.
- Physiology: knowledge of the structure of the human vocal tract, the characteristics of the ear, and most recently, the high-level processing of acoustics and language in the brain.
- Statistics and decision theory: the procedures for comparing patterns based on features and the statistical methods for estimating those features based on the observed signal.
- Information theory and computer science: the set of algorithms and procedures used in pattern matching, including dynamic programming.
- Linguistics: the knowledge of the structure of language, especially of the fundamental phonemic units of speech, and how they are used in speech production. Computational linguists have recently developed models of language, particularly word and concept associations in speech.
- Applied psychology: the human factors of speech recognition applications, and knowledge of the speech production and perception techniques used by human beings.
- Digital signal processing: efficient numerical techniques for analysis of a time-varying signal, especially digital filtering and linear predictive coding. Signal enhancement is also important if noise or channel distortion is present.
- Silicon technology: advances in VLSI silicon processing have been responsible for orders of magnitude increases in processing power and rapidly decreasing costs for speech processing systems.

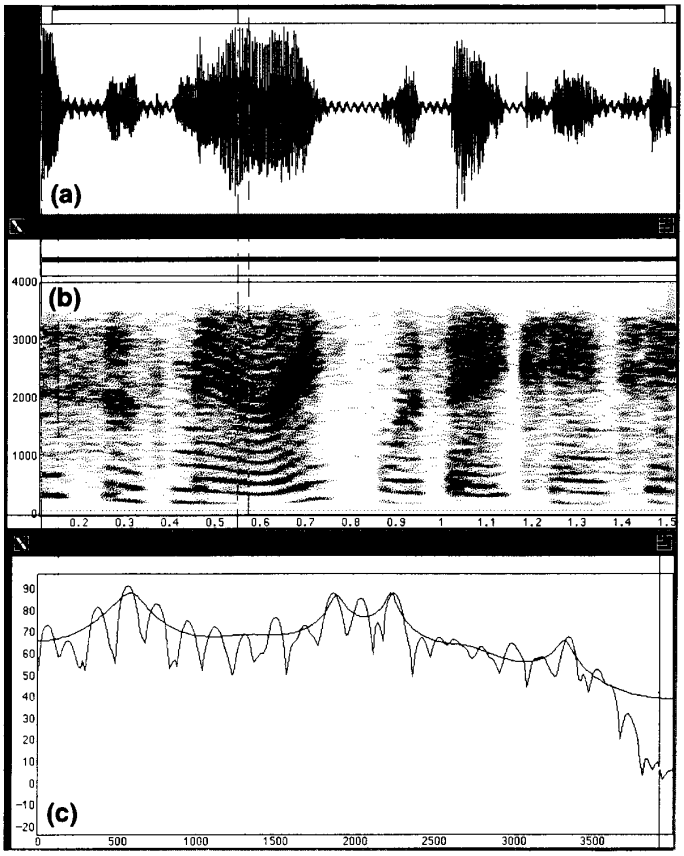
The State of the Art

Dimensions of the Speech Recognition Task

Humans are able to understand speech so easily that we often fail to appreciate the difficulties that this task poses for machines. Here are some of the dimensions in which machine performance falls short:

- Degree of speaker independence: it is easier to characterize an individual speaker's voice than to recognize all voice types and all dialects.

The fundamental difficulty of speech recognition is that the speech signal is highly variable.



■ **Figure 2.** *Analysis of speech for feature extraction. Figure 2(a) is the speech waveform, amplitude vs. time. Figure 2(b) shows the narrow-band spectrogram of the waveform above. A 45 ms speech frame is demarked by the dashed lines. Figure 2(c) shows the power spectrum (bumpy curve) for the marked frame. The smooth curve shows the LPC-smoothed spectrum. Formant peaks at 600 Hz, 1900 Hz and 2300 Hz are evident.*

- Vocabulary complexity: other things being equal, a larger vocabulary is more likely to contain confusable words or phrases that can lead to recognition errors.
- Speaking rate, coarticulation: speech sounds are strongly affected by surrounding sounds in rapid speech. Isolated words are more consistently recognized than words in fluent speech.
- Speaker variability, stress: people are able to understand normal variations in loudness or speed, disregard extraneous coughs or "um"s, and compensate for any involuntary changes caused by stress on the speaker. Only in the simplest cases can machines handle such conditions.
- Channel conditions: poor-quality speech (speech distorted or obscured by noise and extraneous speech) is more difficult for machines to recognize than high-quality speech.

Each of these dimensions of difficulty embodies some aspect of speech variability, which is the central problem of speech recognition. These sources of variability must be carefully considered when planning applications of the technology, because it is these characteristics of robustness that determine whether a speech recognizer will be accurate enough to be satisfactory to the users.

In the pattern matching philosophy of Fig. 1, there are three stages of automatic speech recognition: speech feature analysis, pattern classification, and language processing. This philosophy has been refined over time to address the dimensions of difficulty we mentioned previously.

Speech Feature Analysis

Feature analysis distills the information necessary for speech recognition from the raw speech signal. Just as important, it should discard irrelevant information such as background noise, channel distortion, speaker characteristics, and manner of speaking. Figure 2a shows a speech waveform from which one might wish to extract features for recognition purposes. Figure 2b shows the spectrogram of that speech waveform. The spectrogram is a time-frequency plot of the energy present in the signal as a function of frequency on the vertical axis and time. The horizontal bands in Fig. 2b correspond to the harmonics of the fundamental frequency or pitch of the voice, which is about 200 Hz in this example.

The features used by speech recognition systems may be as simple as the energy or zero-crossing rate of the waveform during each speech frame. A more elegant and robust method for feature extraction is based on the source/filter model of the vocal tract [5]. Linear predictive coding (LPC) analysis calculates the filter characteristics of the vocal tract, and in particular the resonant frequencies, or formants. Characteristics of the source, such as fundamental pitch or absolute energy, are often discarded. Figure 2c shows the power spectrum of the speech in the 45-millisecond frame between the dashed lines in Figs. 2a and 2b. The narrow peaks in the power spectrum are due to harmonics of the fundamental pitch of the voices. The LPC spectrum (the smooth curve in Fig. 2c) exhibits the formants, or resonant frequencies of the vocal tract, that are important to human perception of speech. Experience has shown that in addition to frame-by-frame features, speech recognizers are more immune to channel distortion and speaker variation if short-term time derivatives of the LPC-based spectrum features are also included in the feature vector.

A contrasting method for feature analysis involves the modeling of the human auditory system instead of the vocal tract [11]. This type of analysis usually begins with a set of overlapping bandpass filters more or less similar to the sensitivity of the cochlear membrane, then includes nonlinear effects that occur in human auditory processing. It is important to study both how the ear receives sound and how acoustic features are used by the auditory system and brain to recognize speech.

Both vocal tract modeling and auditory modeling have proven successful in speech recognition. Every improvement in speech feature analysis has resulted in a significant increase in the accuracy of recognition systems. Even though the models used for speech analysis are grounded on solid scientific studies over the past 25 years, current models are oversimplified and represent only the superficial aspects of the physiology of hearing and speech production. Future improvements in speech and hearing models should pay off directly in higher acoustic discrimination power for speech recognizers.

Pattern Classification

The second step of the speech recognition process is pattern classification. There have been four basic classes of pattern matchers for speech recognition over the past 25 years: template matchers, rule-based systems, neural networks, and Hidden Markov Model (HMM) systems. In each case, the pattern matcher must align a sequence of feature vectors with the optimum sequence of speech units (words or phonemes), taking into account both the variability of speaking rates and the constraints of correct language. Mariani gives a good overview of each of these four philosophies [12].

The central idea of template matching is to store examples of the speech patterns called templates that consist of sequences of feature vectors for each speech pattern. Unknown speech features are compared to each of the templates to find the closest match. Because the rate of speaking may vary, a dynamic time warping technique is used to stretch or shrink the time axis to minimize the distortion to the template.

Rule-based systems take a different approach: they set up a series of criteria in a decision tree to determine which of the units of language is present in the speech signal. For large complex speech recognition tasks, it has proven difficult to create sets of rules that generalize well across the many variations in the speech signal. Another issue is that it is hard to provide error recovery if an incorrect conclusion is drawn at an early decision point. The differences between the rule-based and template approaches resulted in a philosophical split in the research community until the early 80s, when both approaches were surpassed by a more powerful theory, Hidden Markov modeling.

Hidden Markov Model (HMM) systems are currently the most successful speech recognition algorithms [13], and are used in virtually all applications except where low cost is the overriding concern. The principal advantage of HMM sys-

Hidden Markov Models

The key assumption of the statistical approach to speech recognition is that speech can be modeled statistically during an automatic process. By examining an ensemble of training speech data, a probabilistic model that characterizes the entire ensemble is created. The resulting model, which represents each speech unit (word or sub-word unit), is more powerful and general than a template.

In the Hidden Markov model (HMM) formalism, speech is assumed to be a two-stage probabilistic process [13]. In the first part of the two-stage process, speech is modeled as a sequence of transitions through states. (Figure 3 shows how these transitions may be a Markov process.) The states are not themselves directly observable, but are manifest by observations, or features. Second, the observations (the features) in any state are not deterministic, but are specified by a probability density function over the space of features. The power and flexibility of the statistical approach derives from this two-stage modeling procedure.

As shown in Fig. 3, a speech model consists of the transition probabilities $a_{i,j}$ between states i and j, and the observation probabilities $b_i(\vec{\mathbf{F}})$ of observing the feature $\vec{\mathbf{F}}$ in state i. (A real Markov model representing speech would have many more states than the simple example of Figure 3.) The Markov model is said to be hidden because one cannot directly observe which state the Markov model is in; one can only observe the features generated by that state. The sequence of states can, however, be inferred from the sequence of features by calculating which of the many Markovian paths through the states of the model is most likely, given the transition probabilities $a_{i,j}$ and the observation probabilities $b_i(\vec{\mathbf{F}})$.

There are three mathematical problems that can be solved in the hidden Markov model approach to statistical pattern matching.

1. Decoding: given a hidden Markov model M and a time sequence of observations $\mathbf{O} = \{ \vec{\mathbf{F}}(t), t = 0, 1, \ldots \}$ generated from M, infer which states of the Markov model were most likely to have been occupied at each time t. There is a well known Viterbi dynamic programming solution to the decoding problem.

2. Classification: given a set of several Markov models M_1, M_2, ..., and an observation sequence $\mathbf{O}$, decide which model M generated the sequence $\mathbf{O}$. This is the essence of the speech recognition problem. First, the Baum-Welsh forward-backward algorithm [13] may be used to find the probability $P(\mathbf{O}|M_k)$ of generating the observation sequence $\mathbf{O}$ from the model M_k. (There is an alternate, approximate estimate using the path from the Viterbi solution to the decoding problem.) Then from the probabilities $P(\mathbf{O}|M_k)$ one can find the relative probabilities $P(M_k|\mathbf{O})$ that the word corresponding to the k^{th} model is present in the observation sequence.

3. Training: given several observation sequences $\mathbf{O}_1, \mathbf{O}_2, \ldots$, create the hidden Markov model $M = \{ a_{i,j}, b_i(\vec{\mathbf{F}}) \}$ that maximizes the probability of generating those observations. The model parameters are found by an iterative procedure known as Baum-Welsh re-estimation. An initial HMM is assumed, and the Baum-Welsh forward-backward algorithm is carried out to find the state occupation probabilities as a function of time. The process of re-estimating the model parameters based on those state occupation probabilities turns out to be equivalent to a steepest-descent gradient search procedure.

There are some theoretical shortcomings of the HMM approach. One of these is the dubious assumption that speech is a strictly Markovian process. As a correction, the state transition probabilities $a_{i,j}$ can be treated as time varying functions, which allows more accurate modeling of the state duration probabilities. Also, the procedure for estimating the HMMs can be modified to minimize the probability of confusing an observation sequence with an incorrect model, as well as to maximize the probability of a particular model. Despite the invalid assumptions of HMMs, this technique has been most the successful in a wide variety of difficult speech-recognition applications.

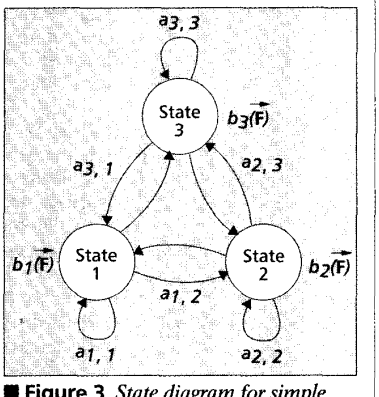


Figure 3. *State diagram for simple Hidden Markov Model. Transition probabilities between states are denoted by arrows. The probability of a transition between state* i *and state* j *is* $a_{i,j}$. *Each state* i *is characterized by a probability density function* $b_i(\vec{\mathbf{F}})$: *the probability of observing acoustic features* $\vec{\mathbf{F}}$ *given that the model is in state* i.

In the laboratory, speech recognizers are quite accurate in acoustic pattern matching. In "real-world" conditions, the error rate is much higher.

tems over template-based approaches is that HMMs retain more statistical information about the speech patterns than templates. While templates consist of one or more exemplars of a multidimensional distribution, HMMs retain information about the complete distribution of features present in the training data. This translates to greater discrimination power (as discussed in the sidebar entitled "Hidden Markov Models"). HMMs use dynamic programming techniques reminiscent of template matching to achieve time normalization. A large part of the improvement in speech recognition systems since the late 1960s is due to the power of this statistical approach, coupled with the advances in computer technology necessary to implement HMMs.

The neural network approach is at least partly motivated by a desire to capture some of the architecture of the human brain, as well as some of its pattern matching capability. This concept was applied to speech recognition in the mid-1980s. In this connectionist approach, pattern classification is done with a multilayer network of perceptrons, with each link between perceptrons assigned a weight determined during a training process. Each perceptron finds the sum of its inputs, partially clips it with a sigmoid function, and passes the result through the links to the next layer of the network. It has proven difficult for neural networks to achieve the level of time alignment of the speech signal that HMMs have attained. So, neural networks are most often used today as static discriminators in systems based on a HMM statistical framework.

Language Processing

In simple applications, the language processing step can be trivial. For isolated word recognition, all that is required is a decision mechanism between alternate words, and a reliability check to insure that the acoustic score of the best word is good enough. This would be sufficient for more complex tasks as well if speech recognizers always made the correct decision based on acoustic information alone. Often there are similar sounding words (or homonyms) for which the grammatical and semantic structure provide disambiguation. For instance, the fragment "one cent" might be confused with "one sent" or even "won scent" judged on acoustic information alone.

The field of language theory includes the study of the acoustic, phonetic, phonological, and semantic structures of speech as well as the mathematics of formal languages. There are two fundamental problems in applying language modeling to speech recognition. First, how can one mathematically describe the structure of a language's valid sequences of words or phonemes; that is, given a putative sequence of symbols, how can one determine the likelihood that the sequence is valid in human language. Second, given such a mathematical description of the language, how can one efficiently compute the optimum sequence of symbols from the acoustic pattern classifier that meets that mathematical description? Both of these questions are exceedingly difficult problems for which linguists lack elegant solutions. In practice, simple language models are used to provide some feedback about the likelihood of words in the language model, however imperfect, to the pattern classifier. Trigram language models, in which each word's probability is evaluated in terms of a triplet of words, have proven effective in voice-dictation systems.

Of the three stages of speech recognition — feature analysis, pattern classification and language modeling — language modeling is the weak link at this time. The difficulties of pattern classification and feature analysis have been solved at least partly by "brute force" techniques aided by fast, powerful computers. But the technology of language modeling is still in its infancy, and has in no way approached the capability of human beings. The fundamental difficulties of language understanding that Pierce pointed out have not yet been solved. Recent research in computational linguistics and statistical language models seems very promising as a scientific base from which to attack these formidable problems [14]. The concept of mutual information between words has recently been generalized from bigram and trigram models to statistical associations based on part-of-speech. There is the potential in the next decade for significant improvement in these areas.

Current Accuracy of Speech Recognition Systems

In the laboratory, speech recognizers are quite accurate in acoustic pattern matching. In "real-world" conditions, the error rate is much higher, due in part to the increased variability of speech styles encountered. Error rate is usually measured as word-substitution errors plus word insertions plus word deletions, divided by the number of words in the speech. Given high-quality, consistent speech samples recorded in a quiet laboratory, and given sufficient speech samples to fully train an HMM, accuracies are within sight of human accuracies in acoustic perception. For instance, numbers can be recognized with an error rate less than one in 300 words (99.7 percent accuracy), for a speech database recorded under laboratory conditions in a sound proof booth [15]. On a larger, speaker-independent task, the DARPA (now known as ARPA, or Advanced Research Projects Agency) resource management task, word accuracies of about 96 percent can be achieved with a vocabulary of a thousand words [16].

In applications, the variability of speech and speaking environments is much greater, so that the same speech recognition algorithm will have error rates much higher than with well controlled laboratory speech. For instance, on connected digits spoken by merchants reading credit card numbers at retail stores, the error rate rises to about 2 percent per digit (from 0.3 percent) with an algorithm [17] similar to that described in reference [15]. This increase in error rate is typical of the difference between laboratory quality speech and the speech encountered in field conditions. Systems engineers who design speech recognition applications must be aware of the increase in recognition errors in the field.

Applications of Speech Recognition

It is important to bear in mind that speech recognition, notwithstanding advances in reliability, remains error-prone. Therefore, the first successful products and services will be those that have the following characteristics:

- Simplicity: successful services will be easy to use.
- Evolutionary growth: the first applications will be extensions of existing systems, such as TouchTone replacement for voice-response systems.
- Tolerance of errors: given that any speech recognizer will make occasional errors, the inconvenience to the user should be minimized. Careful design of human factors will be essential in providing suitable systems.

The central questions when considering an application using a speech recognizer are: 1) what accuracy will the user of this service expect?; 2) is the speech recognizer accurate enough to meet that expectation of the user?; and 3) does the benefit of using speech recognition in this application outweigh its cost, compared to alternative technologies?

Telecommunications and Speech Recognition

Speech recognition is an enabling technology that allows people to interact with computers over telephone lines. Two classes of applications are beginning to appear. The first, cost reduction applications, are tasks in which a person is currently trying to accomplish a task by talking with a human attendant. In such applications, the accuracy and efficiency of the computer system that replaces the attendant is of paramount concern. This is because the benefits of ASR technology generally reside with the corporation that is reducing its costs, and not necessarily with the end users. Hence, users may not be sympathetic to technology failures. Examples of such applications include:

- Automation of operator services, currently being deployed by AT&T and Northern Telecom
- Automation of directory assistance by NYNEX and Northern Telecom
- Control of network fraud by Sprint and AT&T

The second class of applications are services that generate new revenues. For these applications the benefit of speech recognition technology generally resides with the end user, hence they will be more tolerant of technological limitations. Examples include:

- Automation of banking services (Nippon Telephone)
- TouchTone and rotary phone replacement (AT&T)
- Stock quotation services (Bell Northern Research, BNR)

The general concept is to provide voice access to information over the telephone. People will get information from computer databases by asking for what they want, not by typing commands at a computer keyboard. As this technology develops, the voice-response industry will expand to include voice access to information such as weather, traffic reports, news, sports scores, and even information about retail stores. Ease of use is the key.

Beginning in 1985, AT&T had begun investigating the possibility of using limited-vocabulary, speaker-independent speech recognition capabilities to automate a portion of calls currently handled by operators. The introduction of such a service would reduce operator workload, while greatly increasing the overall efficiency for operator handled calls. The exact task studied was the automation of the billing functions — collect, calling card, person-to-person, and bill-to-third-party. Customers would be asked to identify verbally the type of call they wished to make without directly speaking to a human operator. Could a simple five word vocabulary (the function names and operator for human assistance) be designed, built and deployed with such a degree of accuracy that customers would be willing to use the technology? Early trials in 1986 and 1987 proved that the technology did provide such performance levels. After extensive field trials in Dallas, Seattle, and Jacksonville, Fla. during 1991 and 1992, AT&T announced that it would begin deploying Voice Recognition Call Processing. This service automates the front end as well as the back end of collect, calling card, person-to-person and bill-to-third-number calls. The trials were considered successful not just from a technology point of view, but also because customers were willing to use the service.

In 1989, BNR began deploying Automated Alternate Billing Services through local telephone companies in the United States, with Ameritech being the first [18]. For this service, ASR technology was used to automate the back-end of collect and bill-to-third-number calls. After the customer places a call, a speech-recognition device is used to recognize the called party's response to the question: "You have a collect call. Please say yes to accept the charges or no to refuse the charges."

The speech recognition technology differentiates the earlier BNR system from the AT&T system. Analysis of the 1985 AT&T trials indicated that about 20 percent of user utterances contained not only the required command word, but also extraneous sounds that ranged from background noise to groups of nonvocabulary words (as in, "I want to make a collect call please."). These extraneous sounds violated a basic assumption for many speech recognition systems: that the speech to be recognized consist solely of words from a pre-defined vocabulary. In 1990, AT&T developed its word-spotting technology — the ability to recognize key words from a vocabulary list spoken in an unconstrained fashion [19]. Field trials have shown that the ability to spot the key words in speech is a prerequisite for most telephone network applications, and that the ability to recognize speech spoken over voice prompts, called barge-in, is important for mass deployment of ASR technology in networks.

Northern Telecom recently announced the testing of the automation of a second operator function, Directory Assistance [20]. This service would rely on a new, more powerful technology called Flexible Vocabulary Recognition. By entering pronunciation of words in phonetic form, pattern-matching methods can be used to find sequences of phonemes that match sequences in the pronunciation "dictionary." Thus, vocabularies of hundred or thousands of words can, in principle, be recognized without having to record each word. The designers record people speaking words containing each of the phonemes in a variety of contexts. This is especially convenient for vocabularies for which new words need to be added when the service is already in use, such as names in a telephone directory.

Subword-based speech recognition has recently been applied to an information access system

Field trials have shown that the ability to spot the key words in speech is a prerequisite for most telephone network applications.

We believe that there is no fundamental principle that prohibits the construction of a "human-like" speech recognizer — the question is how *to do so.*

by Bell Northern Research working with Northern Telecom [20]. Using the 2000 company names on the New York Stock Exchange, entered in phonetic form, callers to this system can obtain the current price of a stock simply by speaking the name of the stock. Though the accuracy is not perfect, it is adequate, considering the possibilities of confusion on such a large vocabulary. The experimental system is freely accessible on an 800 number, and it has been heavily used since its inception in mid-1992. There are thousands of calls to the system each day, and evidence suggests that almost all callers are able to obtain the quotation they want.

Voice Dictation

The concept of a voice-dictation machine has been an inspiration for much speech research. As predicted by Pierce, the system that will perfectly transcribe naturally spoken spontaneous speech dictated by any speaker is still far from reality. However, there are systems today that work surprisingly well, considering the technical obstacles: Dragon Systems' 30,000-word dictation system, offered in 1990; Kurzweil AI's 50,000-word system for medical applications; and IBM's 20,000-word system based on their speech recognition research, announced in 1992 [21].

The best of the currently available systems work as follows: The prospective user must first train the system to his or her voice, either by speaking a prescribed series of sentences, or by letting the system adapt to his or her voice during an initial period of dictation. All sentences must be spoken with words separated by brief pauses. Just as typing is a skill that must be practiced, dictating by voice with these systems requires an initial effort, both in learning the style of speaking, in planning a sentence ahead of time to avoid false starts and mis-speaking, and in learning how to correct and edit the errors that inevitably occur. Speaking rates of more than 50 words per minute can be achieved after some practice. As the user dictates sentences in this style to the machine, the words appear on the screen. If an error is made, the user can go back and select alternate candidates the system suggests, or repeat the word in question. The word-error rate depends on the skill of the speaker and the similarity of the text to the language model. Word-error rates as low as 3 to 5 percent are possible. These systems employ a model of English, including word frequency, word-pair frequency, and word-trigram frequency to infer which of several similar-sounding words was spoken.

Personal computers that accept voice commands are now available with vocabularies of hundreds of words. Reliable speech recognition accuracy is required to make voice control of computers a legitimate alternative to the keyboard or the mouse.

Speech Understanding for Data Retrieval

Since 1986, ARPA has been funding a large research effort in speech and natural language processing to ensure the availability of these technologies as needed by the United States government. The ARPA Spoken Language program has focused on two main areas. The first is large-vocabulary, continuous-speech recognition. The second, spoken-language understanding, is aimed at the class of interactive problem-solving applications. Both efforts aim to provide real-time, speaker-independent or speaker-adaptive speech recognition technology to handle spontaneous, goal-directed, natural-language speech. ARPA strives for a highly synergistic research program with emphasis on regular formal performance evaluations to track technology advances. The ARPA effort includes many of the major speech research laboratories in the United States: AT&T, BBN, Brown University, Boston University, Carnegie-Mellon University, Dragon Systems, IBM, Massachusetts Institute of Technology (MIT) and MIT Lincoln Laboratories, Paramax, Stanford Research Institute, and Texas Instruments.

From 1987 to 1992, performance evaluations for large vocabulary speech recognition were focused on the 1000-word Resource Management corpus, which consists of read queries and commands. During the five years that this common database was used for testing research advances, speaker-independent accuracy improved from about a word accuracy of 45 percent to 80 percent, using no grammatical information, and from 79 percent to 96 percent using a word-bigram grammar [16].

In 1991, spoken-language understanding research was begun with the collection of spontaneous queries about air travel [16]. The Air Travel Information System (ATIS) corpus was collected from subjects interacting with a simulated understanding system (a so called "wizard" system) that contained data from the Official Airline Guide. Although the ATIS task is much harder than the Resource Management task — it must deal with spontaneous speech, out-of-vocabulary words, and varied channel and microphone conditions — system performance has remained high. Several groups have built on-line demonstrations of this task that run in real time on a workstation. Although not ready yet for field applications, the use of speech recognition technology in highly constrained tasks such as ATIS will be within the range of practicality within the next few years. Deployment will be especially attractive in applications in which information can be obtained remotely by voice.

Speech Recognition in Consumer Products

Speech recognition is being applied to some consumer products such as telephone sets, car stereos, and games. Mobile telephones for automobiles include voice dialing as a safety feature. The call can be made by speaking a name which has been previously trained, by saying a designated key word (such as "office" or "redial"), or by speaking a sequence of isolated digits of the phone number. Manufacturers such as AT&T, Novatel, NEC, Oki, Motorola, and Audiovox have such telephone sets, each with slightly different strategies for voice dialing. Car stereo systems may also be voice controlled. Blaupunct and Clarion have introduced products with this capability. Two companies, Matsushita and Voice Powered Technology, have incorporated voice recognition into video cassette recorders, so that it is possible to program the time and duration of recording entirely by voice. Speech-recognition chips from Texas Instruments that recognize "yes" and "no" have also been used in educational toys and dolls since the mid-1980s.

For most consumer applications of speech recog-

nition, low cost is paramount. The processors in such products typically run algorithms with low computational complexity and limited memory requirements. The successful applications tend to be those that can tolerate a moderate error rate without inconvenience to the consumer.

Prognosis

It has been observed that predictions of future technologies tend to be overly optimistic for the short term and overly pessimistic for the long haul. Such forecasts can have the unfortunate effect of creating unrealistic expectations, followed by premature abandonment of the effort. This article attempts to avoid this effect by carefully pointing out the limitations of speech recognition. We believe that there is no fundamental principle that prohibits the construction of a "human-like" speech recognizer. The question is only *how* to do so.

As Yogi Berra observed, "It's hard to make predictions, especially about the future." Predicting 25 years in the future may be futile because it is impossible to predict when a revolution will occur; few people could have predicted in the 1960's the impact that VLSI would have on our society. All we can say with assurance is that the present course of our technology will take us somewhat further; there are still engineering improvements that can be built on today's science. We can anticipate advances in scientific knowledge that will create a base upon which a new generation of speech recognizers will be designed.

The ultimate long-range goal is to converse fluently with a computer. Voice interaction with a computer has two advantages: first, speech is the most natural means of human communication. Speech recognition puts the burden on the machine to accommodate human skill in speaking and listening, rather than imposing on a person to communicate in a way convenient to the machine (such as with a keyboard). Second, remote access to computers over the telephone is possible because no keyboard or display is needed. Indeed, the desire for remote access to information is the driving force behind the success of voice response systems. These motivations will continue to be powerful stimulants for further research in speech recognition. Some specific predictions:

- Algorithms for topic-specific, speaker-independent recognition of large vocabularies will soon become available. Before the year 2000, this technology will be accurate enough to be successfully used in specific, highly structured applications.
- Major advances will be made in language modeling for use in conjunction with speech recognition. In contrast to the past two decades, in which advances were made in feature analysis and pattern comparison, the coming decade will be the period in which computational linguistics makes a defining contribution to "natural" voice interactions. The first manifestations of these better language models will be in restricted-domain applications for which specific semantic information is available.
- Despite the advances in language modeling, the speech understanding capability of computers will remain far short of human capabilities until well into the next century. Applications that depend on language understanding for unrestricted domains will remain a formidable challenge, and will not be successfully deployed until fundamental advances in understanding of the structures of spoken language.
- Speech recognition over telephone connections will continue to be the most important market segment of speech recognition, both in terms of the number of users of this technology and in terms of its economic effects. The ability to get information remotely will drive many applications.
- "Simple" applications of speech recognition will become commonplace. By the year 2000, more people will get remote information via voice dialogues than by typing commands on computer keyboards to access remote databases. These applications will begin as highly structured dialogues, and will be specific to narrow domains such as weather information or directory assistance.
- Voice command and control will be available as a software option on most computers by 1996. These voice-command options will be used in niche applications. The majority of computer users will bypass this technology in favor of the keyboard and the mouse.
- People will learn to modify their speech habits to use speech recognition devices, just as they have changed their speaking behavior to leave messages on answering machines. Even though they will learn how to use this technology, people will always complain about speech recognizers.
- Finally, we confidently predict that at least one of these eight predictions will turn out to have been incorrect.

Over the next 25 years, will speech-recognition technology result in a natural language Turing machine that is indistinguishable from a human being, or in creating the robots of science-fiction that can speak, listen, and understand? The answer is certainly "no." In this sense, Pierce was right in his prediction: Computers still lack the ability to make elementary inferences based on human experience. Instead, speech recognition will proceed incrementally, but inevitably, forward. As the science and technology advance together, applications will be deployed — applications that seem simple, but that prove beneficial to society. Sooner than we might expect, applications based on speech recognition technology will touch the lives of every one of us.

The speech understanding capability of computers will remain far short of human capabilities until well into the next century.

References

[1] J. R. Pierce, "Whither Speech Recognition?", *JASA*, vol. 46, no. 4, pp. 1029-1051 (1969).
[2] G. R. Doddington, "Whither Speech Recognition?", in Trends in Speech Recognition, W. Lea, ed., (Prentice-Hall, 1980).
[3] There are several newsletters covering the industry, including *Speech Technology: Man/Machine Voice Communications*, Media Dimensions Inc., New York; *Speech Recognition UPDATE*, W. Meisel, ed., TMA Associates, Encino, Calif.; and *ASR News*, Voice Information Associates Inc., Lexington, Mass.
[4] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, (Prentice-Hall, 1993).
[5] J. L., Flanagan, Speech Analysis, Synthesis and Perception, 2nd ed., (Springer-Verlag, New York, 1972).
[6] D. O'Shaughnessy, Speech Communication: Human and Machine, (Addison-Wesley, 1987).
[7] N. Jayant, "Signal Coding: Technology Targets and Research Directions", *IEEE JSAC*, vol. 10, no. 5, pp. 830-849, June 1992.
[8] R. Carlson, "Models of Speech Synthesis", in Human/Machine Communication by Voice, in D. B. Roe and J. G. Wilpon, eds., (National Academy of Sciences Press, to be published).
[9] D. B. Roe and J. G. Wilpon, eds., Human/Machine Communication by Voice, (National Academy of Sciences Press, to be published).
[10] S. E. Levinson and D. B. Roe, "A Perspective on Speech Recognition", *IEEE Commun. Mag.*, vol. 28, no. 1, pp. 28-34, Jan. 1990.
[11] J. B. Allen, "Micromechanical Models of the Cochlea", *Physics Today*, pp. 40-47, July 1992.
[12] J. Mariani, "Recent Advances in Speech Processing", *Proc. IEEE*

The ultimate long-range goal is to converse fluently with a computer.

Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 89, vol. S1, pp. 429-440, 1989.

[13] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[14] M. Marcus, "New Trends in Natural Language Processing", in *Human/Machine Communication by Voice*, D. B. Roe and J. G. Wilpon, eds., (National Academy of Sciences Press, 1994).

[15] J. Gauvin and C. H. Lee, "Improved Acoustic Modeling with Bayesian Learning," *Proc. IEEE ICASSP 92*, pp. 481-484, 1992.

[16] M. Marcus, ed., *Proc. of the Fifth DARPA Speech and Natural Language Workshop*, San Mateo, Calif., Morgan Kaufmann Publishers, 1992.

[17] P. Ramesh *et. al.*, "Speaker Independent Recognition of Spontaneously Spoken Connected Digits," *Speech Communication*, vol. 11, pp. 229-235, 1992.

[18] M. Lennig, "Putting Speech Recognition to Work in the Telephone Network," *Computer*, vol. 23, no. 8, pp. 35-41, Aug. 1990.

[19] J. G. Wilpon *et. al.*, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol. 38, no. 11, pp. 1870-1878, Nov. 1990.

[20] M. Lennig *et al.* "Flexible Vocabulary Recognition of Speech", *Proc. ICSLP-92*, pp. 93-96, Banff, Canada, Oct. 1992; M. Lennig *et. al.*, "Automated Bilingual Directory Assistance Trial in Bell Canada," *Proceedings of the 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Piscataway, N.J., Oct. 1992.

[21] L. R. Bahl *et. al.*, "Large Vocabulary Natural Language Continuous Speech Recognition," *Proc. IEEE ICASSP-89*, pp. 465-468, May 1989.

Biographies

David B. Roe received a Ph.D in experimental low-temperature physics from Duke University in 1976 and joined Bell Laboratories, in 1977. In 1986 he became a member of the Speech Research Department in Murray Hill, New Jersey, where he is now a supervisor. His current research interests are speech recognition, spoken-language translation, DSP implementations of speech-recognition algorithms, and application of speech technologies to telecommunications.

Jay G. Wilpon received B.S. and A.B. degrees in mathematics and economics, respectively, from Lafayette College in 1977. He obtained an M.S. degree in electrical engineering/computer science from Stevens Institute of Technology in 1982. Since 1977, he has been with the Speech Research Department at AT&T Bell Laboratories, Murray Hill, New Jersey, where he is a distinguished member of the technical staff. He has been engaged in speech communications research and is presently concentrating on problems in automatic speech recognition. He has published extensively in this field and has been awarded several patents. His current interests lie in keyword spotting techniques, large vocabulary spoken language understanding systems, speech recognition training procedures, and determining the viability of implementing speech recognition systems for general usage over the telephone network. He is vice-chairman of the IEEE Digital Signal Processing Society's Speech Committee and a member of the ARPA Spoken Language Understanding Coordinating committee. In 1987 he received the IEEE Acoustics, Speech and Signal Processing Society's Paper Award for his work on clustering algorithms for use in training automatic speech recognition systems.